THE CENTRE FOR
LONG-TERM RESILIENCE

January 2025

Biological Tools and the EU AI Act

Report

Richard Moulange, Tina Wünn, Cassidy Nelson*

*If you have any queries about this report, please get in touch at biosecurity@longtermresilience.org

Executive Summary

Context

The European Union AI Act (EU AI Act) is a regulatory framework that aims to govern the development, deployment, and use of AI systems within the EU. It focuses on large-scale AI systems classified as *general-purpose AI* (GPAI). The Code of Practice accompanying the EU AI Act is currently being drafted and will detail the rules for GPAI providers of models deemed to pose *systemic risks*.

While GPAIs are often large language models (FLI, 2022), other non-natural language AI systems have become increasingly common across various domains. This includes numerous *AI-enabled biological tools*[1] which have emerged in life sciences research and industry. Such tools include both *narrow* AI models that excel at one particular task, such as protein design or toxicity prediction, and *biological foundation models* (bioFMs)—larger-scale AI systems trained on biological data such as nucleic or amino acid sequences. Previous work by the Centre for Long-Term Resilience (Rose & Nelson, 2023; Moulange et al., 2024) has defined and categorised AI-enabled biological tools and examined their misuse-relevant capabilities.

It is plausible that biological AI models could pose reasonably foreseeable negative effects to public health, safety or public security at scale: such models could potentially meaningfully assist malicious actors with the development or acquisition of biological weapons.[2] The empirical evidence base for determining the level of risk AI models pose is still under development. However, it was recently demonstrated that protein design tools could redesign sequences to evade nucleic acid synthesis screening (Wittmann et al., 2024).

It is currently undetermined whether AI-enabled biological tools are subject to regulation under the EU AI Act. Biological AI models would be classified as GPAIs if they are considered to meet certain generality and downstream integration criteria and are not solely used for research and development purposes (the *R&D exemption*). GPAIs with *high-impact capabilities* and reasonably foreseeable negative effects are deemed to pose systemic risks. The second draft of the Code of Practice states that biological weapon threats (including risks of 'significantly lowering the barriers of entry for malicious actors in the development, design, acquisition, and use of weapons') should be treated as systemic risks.

In this brief, we apply the relevant GPAI and systemic risk definitions in the EU AI Act to 50 AI biological models and examine the implications.

[1] AI-enabled biological tools are AI models used for life sciences tasks trained on biological data using machine learning techniques, such as deep neural networks. They can include biological design tools such as those used for protein or vaccine design, and other non-design tools such as experimental simulation platforms.

[2] This is recognised explicitly in the EU AI Act (Recital 110) and was also highlighted at the Seoul AI Summit. It has been explored in a number of different ways by the Centre for Long-Term Resilience (see Rose & Nelson, 2023, Moulange et al., 2024, Smith et al., 2024 and Rose et al., 2024, for example).

Summary Findings

1. We assess that *ESM C 6B* and *ESM3*—probably currently the most advanced protein language and generative bioFMs available, respectively—*almost certainly*[3] meet the generality and downstream integration GPAI criteria and *likely* have been made available commercially in the Union market, negating the R&D exemption. Therefore, we assess overall it is *likely* that *ESM C 6B* and *ESM3* are GPAI models.

2. We assess that six additional bioFMs (*Evo, EvoDiff, Chai-1, xTrimoPGLM-100B, ProtT5-XL-UniRef50* and *Nucleotide Transformer) likely* meet the generality and downstream integration criteria in the GPAI definition, but only four (*Evo, EvoDiff, Chai-1, ProtT5-XL-UniRef50)* would be classed as GPAI since the remaining two are *likely* to fall under the R&D exemption.

3. Of the non-bioFM models that we considered, we found three examples of experimental design, planning, and simulation tools (*AutoOED, Synthace, Benchling for Lab Automation*) and two examples of autonomous experimental platforms (*Eve, BioAutomata*)—where there was a *realistic possibility* that these models meet the generality and integration criteria for GPAI. These AI systems are commercially available, however, we determined it was *unlikely* that they met the systemic risk criteria.

4. On examination of the above potential biological GPAIs, we determined it is a *realistic possibility* that *ESM3* and *ESM* C 6B meet the high-impact capabilities threshold for protein language and protein generation models, respectively, and therefore pose systemic risk as defined by the EU AI Act.

5. We think that it is *highly likely* that some future biological GPAIs will be classified as posing systemic risk given that ESM3 was trained with 10^24 FLOPs, close to the 10^25 FLOPs threshold specified in Chapter 5, Article 51. Future models exceeding this will be automatically considered to possess high-impact capabilities and pose systemic risk.

Brief Definitions

This section briefly describes relevant terms in the EU AI Act in order to contextualise our main findings. The full definitions and related recitals are available in Table 2 at the end of this report.

'General-purpose' criteria, with R&D exemption:

The GPAI definition in Chapter 1, Article 3 (63) requires a model to meet two generality criteria and one downstream integration criterion:

(a) '[display] significant generality' and 'competently [perform] a wide range of distinct tasks'

[3] We use the terminology of the Professional Head of Intelligence Assessment's Probability Yardstick to specify our level of certainty for our conclusions. This yardstick splits the probability scale into seven ranges: *remote chance; highly unlikely; unlikely; realistic possibility; likely; highly likely; almost certain.*

(b) be able to be 'integrated into a variety of downstream systems'

Additionally, Recital 98—which, like all recitals, is advisory in nature but not legally binding—clarifies that a model with 'at least a billion parameters' that was 'trained with a large amount of data using self-supervision' should be considered to meet the first two generality criteria. The GPAI definition also includes an *R&D exemption*: models used for 'research, development or prototyping activities before they are placed on the market' are explicitly not considered to be 'general-purpose'.

Classification of GPAI with Systemic Risk:

In Chapter 5, Article 51, a GPAI is classified as having systemic risk if it meets any of the following:

(a) it has high-impact capabilities evaluated on the basis of appropriate technical tools and methodologies, including indicators and benchmarks; or
(b) a decision of 'the Commission, ex officio or … the scientific panel' decides positively that a GPAI has high-impact capabilities; or
(c) the GPAI was trained with more than 10(^25) floating point operations per second (FLOPS) of compute.

Chapter 1, Article 3 (64) clarifies that 'high-impact capabilities' means capabilities that match or exceed the capabilities recorded in the most advanced GPAIs.

Annex XIII, which is legally binding and used to determine if (a) above is met, specifies that the Commission should take into account several criteria to designate a GPAI as having systemic risk, including the number of parameters, the input and output modalities of the model, and the specific type of inputs and outputs (e.g. biological sequences).

Recital 110 accompanying these criteria highlights deliberate biological misuse risks and lowering barriers to biological weapon development and acquisition. Recital 111 highlights the need for a methodology to be established to classify GPAIs with systemic risks.

'Systemic risk':

Chapter 1, Article 3 (65) defines systemic risk to mean a risk that is specific to the high-impact capabilities of general-purpose AI models (see above).

In addition, the model must have a significant impact on the Union market either due to their i) reach or ii) actual or reasonably foreseeable negative effects on public health, safety, public security, fundamental rights, or the society as a whole, that can be propagated at scale across the value chain.

'Placing' and 'Making available' on the market:

(a) Placing on the market: 'the first instance of making an AI system available on the Union market'
(b) Making available on the market: 'the supply of an AI system for distribution or use within the Union market as part of a commercial activity, whether paid or free'

As above, the definition of GPAI excludes AI models used solely for research, development, or prototyping and that have not yet been placed on the market. Once placed on the market, models must comply with the EU AI Act.

Methodology

We modified the categories of narrow AI-enabled biological tools that we previously described (Rose & Nelson, 2023) before examining models in the following categories: Protein design tools; protein structural prediction or representation tools; small biomolecule design tools; vaccine design tools; viral vector design tools; genetic modification tools; genome assembly tools; toxicity prediction or detection tools; pathogen property prediction tools; host-pathogen interaction prediction tools; immunological system modelling tools; experimental design, planning tools and simulation tools; autonomous experimental platforms; and biological foundation models (bioFMs). For each of these 14 categories, we selected between 2-7 AI models.[4] For each model, we examined whether they would fulfil the generality and downstream integration criteria for GPAI classification, specifying our estimates using the Professional Head of Intelligence Assessment's Probability Yardstick. We also considered whether any model was trained with at least 1B parameters on large datasets using self-supervision, drawing on the Epoch AI dataset of biological sequence models. The information we analysed included the underlying academic paper or technical report, and sometimes GitHub documentation or similar.

Findings

GPAI Models

We assess that *ESM C 6B*—probably the most advanced protein representation bioFM—and *ESM3*—the most capable generative protein bioFM currently available—*likely* meet the legally binding definition of GPAI (Chapter 1, Article 3 (63)). *ESM3* and *ESMC 6B* are both capable of performing a wide range of protein design tasks, including generating novel proteins, designing antibodies, self-improving designs based on feedback and targeting designs according to the function requested by the user. These attributes *almost certainly* fulfil the generality criteria. They both also exceed the advisory 1B parameter threshold for generality (Recital 98). The ESM3 and ESM C GitHub documentation explicitly discusses 'standard commercial use like developing molecules and developing downstream ML models and methods' which evidences that they both *almost certainly* meet the downstream integration criterion.

However, it is not clear whether *ESM3* is available within the Union via a commercial licence: ESM3 is certainly available commercially, but not necessarily within the Union.[5] All three *ESM*

[4] We have redacted the full list of models for other categories as part of our commitment to responsible disclosure. For any inquiries, please contact biosecurity@longtermresilience.org

[5] 'ESM3 models remain available for commercial use through partnerships to advance AI for drug discovery and other specific applications' (Evolutionary Scale, 2024)

C models (*300M*, *600M* and *6B*) have been made available commercially via AWS Sagemaker. We therefore assess it is *likely* that *ESM3* is also available commercially within the Union. If not, this will likely happen in 2025 since their developer Evolutionary Scale intends to additionally release commercial versions of both model classes via Omics platform, AWS Bedrock and NVIDIA BioNeMo.

We additionally assessed six other bioFMs: *Evo, EvoDiff, Chai-1, xTrimoPGLM-100B, ProtT5-XL-UniRef50* and *Nucleotide Transformer.* These bioFMs *likely* meet the GPAI requirements, however, two may have the R&D exemption. Each of these models has more than 1B parameters, performs many different protein structural prediction and design tasks, and was trained on a large dataset in a self-supervised way. We expect that they therefore likely meet the generality criteria and—as with other bioFMs—can straightforwardly be integrated with other machine learning systems downstream. None of the models appear to be explicitly marketed as a commercial product; however, *Evo, EvoDiff, Chai-1* and *ProT5-Xl-UniRef50* are all released under licenses that permit commercial use. Since the models' weights can be downloaded in the Union, we conclude they *likely* have been placed on the market—since the Act explicitly includes 'for distribution or use on the Union market in the course of a commercial activity, whether in return for payment or free of charge'—and thus meet the GPAI definition. Note that given the licences for *xTrimoPGLM-100B* and *Nucleotide Transformer* preclude commercial use, these models are *unlikely to meet* the GPAI definition.

Additionally, we found that there is a *realistic possibility* that several experimental, design and planning tools and autonomous experimental tools meet the legally binding GPAI definition. We examined five models (*AutoOED, Synthace, Benchling for Lab Automation, Eve, BioAutomata*) that plausibly meet the generality and multitask threshold. These tools are commercial in nature, which would invalidate the R&D exemption.

For all remaining tools we examined across the remaining 11 categories, we assess that it is *highly unlikely* that any that we assessed meet the GPAI definition. This is both because most of these models are not capable of 'performing a wide range of distinct tasks' and because even the largest (e.g. *AlphaFold3* and *RFdiffusion*) have significantly fewer than one billion parameters. We note that some narrower AI tools outperform bioFMs on specific tasks (for example, *AlphaFold 3* outperforms ESM3 in protein folding) but this does not negate the conclusion they are *highly unlikely* to be considered a GPAI.

Systemic Risk

Annex XIII specifies that the input and output modalities of a model should be taken into account to determine high-impact capabilities for each modality, *ESM3* and *ESM* C 6B it is a *realistic possibility* that they meet the high-impact capabilities threshold for protein language and protein generation models, respectively, and therefore pose systemic risks as defined by the EU AI Act. We assess that it is *unlikely* that other bioFMs would meet the threshold since they appear to have worse performance compared to *ESM3* and likely *ESM C 6B*, however, this cannot be confirmed without extensive technical evaluation beyond the scope of this report. In particular, it may be that another bioFM could be considered the "most advanced"

on some particular biological subtask, even if it does not outperform other models across all biological tasks.

Synthace has previously been directly connected to large language models like ChatGPT-3.5, which have many billions of parameters. This combined system, if itself made commercially available, could exceed the 1B parameter threshold for generality. However, given the models themselves have less than 1B parameters and do not inherently facilitate activities that could lead to significant negative effects on public health, safety, or public security, we deemed that it was *unlikely* that they have high-impact capabilities or pose systemic risks.

None of the models we reviewed, including the bioFMs, were trained with 10(^25) or more floating point operations per second (FLOPS). However, ESM3 was trained with over 10(^24) FLOPS, indicating that this threshold may be passed in the near future. Future AI systems meeting the GPAI definition trained with at least 10(^25) FLOPS would automatically be considered to present systemic risks.

Our findings for models meeting the GPAI definition and parameter count are summarised in Table 1 below.

Table 1. Biological models meeting the GPAI definition or parameter count*

Category	**Definition**	**Models**	**Meets GPAI definition (as per Chapter 1, Article 3 (63))**	**> 1 billion parameters (as per Recital 98)**
Experimental design / planning tools and simulation tools	Tools that are able to generate designs for experiments, given a predefined 'campaign objective' or Tools that are able to simulate (in silico) and predict experimental outcomes	AutoOED Synthace Benchling for Lab Automation	*Realistic possibility* These tools can be used for assay development, media optimisation, and purification process development.	*Remote chance* However, that *Synthace*—one of the tools we assessed—has also been connected to ChatGPT-3.5 in the past, which could potentially form a general-purpose AI system which would exceed the 1B parameter threshold.
Autonomous experimental platforms	Tools that are able to conduct experiments (including physical tests, modelling or data mining) without human intervention	Eve BioAutomata	*Realistic possibility* These systems perform multiple kinds of experimental tasks.	*Remote chance* We were unable to find parameter count information for the tools we assessed here. However, given that the examples we found were published before 2020 and use traditional machine learning systems, we consider it to be a remote chance that they exceed the 1B parameter threshold.

Biological foundation models (BioFMs)		ESM3 ESM C Chai-1 Evo EvoDiff ProtT5-XL-UniRef50 Nucleotide Transformer xTrimoPGLM-100B	*Almost certainly* The descriptions of these models include mention that they can be fine-tuned / adapted for various tasks and prediction purposes.	*Almost certainly* ESM3: 98B ESM C 6B: 6B Chai-1: Unknown Evo: 7B EvoDiff: 0.64B ProtT5-XL-UniRef50: 3B Nucleotide Transformer: 2.5B xTrimoPGLM-100B: 100B

* We have redacted the full list of models we asssessed for other categories as part of our commitment to responsible disclosure. For any inquiries, please contact biosecurity@longtermresilience.org

Limitations

We were unable to fully evaluate whether models with GPAI potential possess high-impact capabilities as set out in Annex XIII since this requires an assessment of the 'appropriate technical tools and methodologies, including indicators and benchmarks'. Such technical evaluations are still very nascent for biological AI models. Furthermore, we are unaware of any technical or regulatory work that unifies the evaluation of advanced biological and natural language AI systems, so it is very difficult to compare a narrow AI-enabled biological tool or bioFM to, for example, a large language model. The absence of robust benchmarking tools for GPAIs, particularly in the biological domain, presents a significant limitation for evaluating systemic risks under the EU AI Act. This gap could lead to inconsistent or subjective assessments of high-impact capabilities. However, we were able to rule out a high-impact capabilities classification based on the 10^25 computational threshold.

This analysis provides a lower-bound assessment given that, while we evaluated over 50 tools, we did not conduct a systematic review of all biological tools. Parameter counts were not publicly available for all models, and the commercial availability of some models in the EU was not always able to be determined. In such instances, we took the conservative approach of assuming such criteria were not met. Given the rapid pace of development in biological AI models, this analysis represents a snapshot in time where capabilities may quickly evolve. The field's dynamic nature means that models' capabilities, commercial status, and potential risks could change significantly over short periods.

Conclusion

Our analysis suggests that it is likely that two currently available biological foundation models—*ESM C 6B* and *ESM3*—meet the EU AI Act's definition of a general-purpose AI system and that it is a realistic possibility that they pose systemic risk based on the high-impact capabilities criteria. Other biological foundation models and narrower AI biological tools we examined are unlikely to meet both the GPAI or systemic risk criteria at present, but future models could automatically qualify if they exceed the 10^25 FLOPs threshold. Our certainty remains bounded by debatable interpretations of high-impact capabilities and the lack of robust benchmark tools for biological AI models. We recommend ongoing assessment of new and emerging biological models, in particular biological foundation models, as future AI systems may meet both GPAI and systemic risk criteria and thus be subject to regulation under the EU AI Act.

Appendix

Table 2. Full EU AI Act definitions

Term / Classification	Definition*	Relevant Recitals†*
General-purpose AI model	Chapter 1, Article 3 (63): 'general-purpose AI system' means an AI model, including where such an AI model is trained with a large amount of data using self-supervision at scale, that **displays significant generality** and is capable of competently **performing a wide range of distinct tasks** regardless of the way the model is placed on the market and that **can be integrated into a variety of downstream systems** or applications, **except AI models that are used for research, development or prototyping activities before they are placed on the market**	Recital 97: [...] This Regulation provides specific rules for general-purpose AI models and for general-purpose AI models that pose systemic risks, which should apply also when these models are integrated or form part of an AI system. [...]
		Recital 98: Whereas the generality of a model could, inter alia, also be determined by a number of parameters, models with **at least a billion of parameters and trained with a large amount of data using self-supervision** at scale should be considered to display significant generality and to competently perform a wide range of distinctive tasks.
		Recital 99: Large generative AI models are a typical example for a general-purpose AI model, given that they allow for flexible generation of content, such as in the form of text, audio, images or video, that can readily accommodate a wide range of distinctive tasks.

General-purpose AI system	Chapter 1, Article 3 (66): 'general-purpose AI system' means an AI system which is based on a general-purpose AI model and which has the capability to serve a variety of purposes, both for direct use as well as for integration in other AI systems;	Recital 100: When a general-purpose AI model is integrated into or forms part of an AI system, this system should be considered to be [a] general-purpose AI system when, due to this integration, this system has the capability to serve a variety of purposes. A general-purpose AI system can be used directly, or it may be integrated into other AI systems.
Systemic risk	Chapter 1, Article 3 (65): 'systemic risk' means a risk that is **specific to the high-impact capabilities of general-purpose AI models**, having a significant impact on the Union market due to their reach, or due to **actual or reasonably foreseeable negative effects on public health, safety, public security**, fundamental rights, or the society as a whole, **that can be propagated at scale across the value chain**	Recital 110: [...] In particular, international approaches have so far identified the need to pay attention to risks from potential intentional misuse or unintended issues of control relating to alignment with human intent; chemical, **biological**, radiological, and nuclear risks, such as the ways in which **barriers to entry can be lowered, including for weapons development, design acquisition, or use**; [...] risk that a particular event could lead to a **chain reaction with considerable negative effects** that could affect up to an entire city, an entire domain activity or an entire community.
High-impact capabilities	Chapter 1, Article 3 (64): 'high-impact capabilities' means capabilities that match or exceed the capabilities recorded in the most advanced general-purpose AI models;	Recital 110 (as above)

Classification of General-Purpose AI Models as General-Purpose AI Models with Systemic Risk	Chapter 5, Article 51: 1. A general-purpose AI model shall be classified as a general-purpose AI model with systemic risk if it meets any of the following conditions: a) it has high impact capabilities evaluated on the basis of appropriate technical tools and methodologies, including indicators and benchmarks; b) based on a decision of the Commission, ex officio or following a qualified alert from the scientific panel, it has capabilities or an impact equivalent to those set out in point (a) having regard to the criteria set out in Annex XIII. 2. A general-purpose AI model shall be presumed to have high impact capabilities pursuant to paragraph 1, point (a), when the cumulative amount of **computation** used for its training measured in floating point operations is greater than **10(^25)**.	Recital 110 (as above) Recital 111: It is appropriate to establish a methodology for the classification of general-purpose AI models as general-purpose AI model with systemic risks. Since systemic risks result from particularly high capabilities, a general-purpose AI model should be considered to present systemic risks if it has high-impact capabilities, evaluated on the basis of appropriate technical tools and methodologies, or significant impact on the internal market due to its reach. High-impact capabilities in general-purpose AI models means capabilities that match or exceed the capabilities recorded in the most advanced general-purpose AI models. The full range of capabilities in a model could be better understood after its placing on the market or when deployers interact with the model. According to the state of the art at the time of entry into force of this Regulation, the cumulative amount of computation used for the training of the general-purpose AI model measured in floating point operations is one of the relevant approximations for model capabilities. The cumulative amount of computation used for training includes the computation used across the activities and methods that are intended to enhance the capabilities of the model prior to deployment, such as pre-training, synthetic data generation and

		finetuning. Therefore, an initial threshold of floating point operations should be set, which, if met by a general-purpose AI model, leads to a presumption that the model is a general-purpose AI model with systemic risks. This threshold should be adjusted over time to reflect technological and industrial changes, such as algorithmic improvements or increased hardware efficiency, and should be supplemented with benchmarks and indicators for model capability. To inform this, the AI Office should engage with the scientific community, industry, civil society and other experts. Thresholds, as well as tools and benchmarks for the assessment of high-impact capabilities, should be strong predictors of generality, its capabilities and associated systemic risk of general-purpose AI models, and could take into account the way the model will be placed on the market or the number of users it may affect. To complement this system, there should be a possibility for the Commission to take individual decisions designating a general-purpose AI model as a general-purpose AI model with systemic risk if it is found that such model has capabilities or an impact equivalent to those captured by the set threshold. That decision should be taken on the basis of an overall assessment of the criteria for the designation of a general-purpose AI model with systemic risk set

		out in an annex to this Regulation, such as quality or size of the training data set, number of business and end users, its input and output modalities, its level of autonomy and scalability, or the tools it has access to. Upon a reasoned request of a provider whose model has been designated as a general-purpose AI model with systemic risk, the Commission should take the request into account and may decide to reassess whether the general-purpose AI model can still be considered to present systemic risks.
Criteria for the Designation of General-Purpose AI models with Systemic Risk Referred to in Article 51	Annex XIII: For the purpose of determining that a general-purpose AI model has capabilities or an impact equivalent to those set out in Article 51(1), point (a), the Commission shall take into account the following criteria: (a) the number of parameters of the model; (b) the quality or size of the data set, for example measured through tokens; (c) the amount of computation used for training the model, measured in floating point operations or indicated by a combination of other variables such as estimated cost of training, estimated time required for the training, or estimated energy consumption for the training; (d) the input and output modalities of the model, such as text to text (large language models), text to image,	

	multi-modality, and the state of the art thresholds for **determining high-impact capabilities for each modality, and the specific type of inputs and outputs (e.g. biological sequences);** (e) the benchmarks and evaluations of capabilities of the model, including considering the number of tasks without additional training, adaptability to learn new, distinct tasks, its level of autonomy and scalability, the tools it has access to; (f) whether it has a high impact on the internal market due to its reach, which shall be presumed when it has been made available to at least 10 000 registered business users established in the Union; (g) the number of registered end-users.	
Placing on the market	Chapter 1, Article 3 (9): 'placing on the market' means the first making available of an AI system or a general-purpose AI model on the Union market	
Making available on the market	Chapter 1, Article 3 (10): 'making available on the market' means the supply of an AI system or a general-purpose AI model for distribution or use on the Union market in the course of a **commercial activity, whether in return for payment or free of charge**	

* Bold emphasis added and does not exist in original text

† In some instances, only portions of recital are included—for the full text, please see the relevant links